Periodica Polytechnica
Civil Engineering

62(2), pp. 318–322, 2018
https://doi.org/10.3311/PPci.10045

K-Nearest Neighbours Method as a Tool for Failure Rate Prediction

Małgorzata Kutyłowska[1*]

RESEARCH ARTICLE

Received 20 September 2016; Revised 23 August 2017; Accepted 02 October 2017

Abstract

The paper shows the results of failure rate prediction using non-parametric regression algorithm K-nearest neighbours. The whole data set for years 1999-2013 was divided randomly into two groups (learning – 75% and testing – 25%). Besides, data from year 2014 were used for verifying the model. The dependent variable (failure rate) was forecasted on the basis of independent variables (number of installed house connections, total length and number of damages of water mains, distribution pipes and house connections). Four types of distance metric: Euclidean, quadratic Euclidean, Manhattan and Czebyszew were checked and four KNN models were created. Taking into consideration all constraints and assumptions, models using Euclidean and quadratic Euclidean distance metrics gave the most optimal prediction results. The optimal number of K nearest neighbours equalled to 2 and 3 concerning models KNN-E, KNN-E2, KNN-C and KNN-M, respectively. Validation error was the smallest for models KNN-E and KNN-E2 and amounted to 0.0130, for model KNN-M was equal to 0.0152 and for KNN-C to 0.0150.

Keywords

failure analysis, K-nearest neighbours, prediction, water-pipe network

1 Faculty of Environmental Engineering
Wrocław University of Science and Technology
Wybrzeże St. Wyspiańskiego 27, 50-370 Wrocław, Poland
* Corresponding author, email: malgorzata.kutylowska@pwr.edu.pl

1 Introduction

Water-pipe networks are one of the most important part of the whole water supply system and belong to the critical infrastructure. The technical condition of the water conduits and amount of water provided to the consumers [1] should be maintained at the proper level concerning forecasting, suitable management and failure analysis. Research devoted to the determination of the water conduit failure intensity indicator and other factors having a bearing on the proper functioning of municipal water networks (e.g. impact of water losses on the soil surrounding the pipe [2]) has been conducted in Poland and in the world for many years [3], [4], [5], [6]. Failure rate of water pipes (λ) should be estimated not only on the basis of operational data, but also using the best available mathematical techniques and models, e.g. [7], [8]. On the other hand, failure analysis could be established using models based on the artificial intelligence [9], [10] or other statistical and probabilistic methods used in water and sewerage systems [11], [12], [13].

Recently, a lot of regression methods (e.g. support vector machine – SVM, regression tress – RT and K-nearest neighbours – KNN) were used to solve many engineering problems. For instance, the localization of leakages from water-pipe network was estimated using SVM algorithm [14], the building damage was assessed by means of RT [15] and KNN algorithm was used relating to time series analysis in industrial processes [16]. The main aim of this paper is to check if non-parametric regression algorithm KNN could be also useful for prediction of indicator λ of water conduits (water mains, distribution pipes and house connections).

2 Material and methods

K-nearest neighbours algorithm is the relatively simple one among other learning methodologies. It is assumed that similar data are grouped to the same class. The prediction is based on the comparison whether forecasted values belong to the exemplary set or not [17]. In the regression problems continuous dependent variable is predicted on the base of independent variables. The choice of the number of *K* nearest neighbours has significant meaning. This parameter is the most important concerning the

prediction quality. The smallest K, the bigger prediction variance. The optimal number of K is not known *a priori* and usage of V-fold-cross-validation algorithm is recommended to find the best K. The main idea of cross-validation method is based on such approach: the data are divided into V (chosen randomly) separate parts. The analysis is carried out for certain values of parameters using V-1 data sets as learning examples. In regression problems the prediction error is calculated as sum of squares of residuals. The procedure is repeated for all V data segments and at the end the errors are averaged. Cross-validation algorithm is related to the estimation of prognostic quality of the model using testing sample which was not known for the model during its creation. In other words, the model is created using learning sample and the real model accuracy is checked using testing sample [17]. After selecting the proper number of K, the prediction could be carried out. In regression problems, the average for K nearest neighbours is calculated according to the equation (1) [17]:

$$y = \frac{1}{K}\sum_{i=1}^{N} y_i \tag{1}$$

where y_i is the output value for i learning example and y is the value of output variable for new example. The result is obtained on the base of the K nearest neighbours of new point. Following this assumption, it is needed to have some kind of measurement of the distance between examples. There are four types of distance metric: Euclidean (E) – equation (2), quadratic Euclidean (E2) – equation (3), Manhattan (M) – equation (4) and Czebyszew (C) – Equation (5) [17]:

$$D(x,p) = \sqrt{(x-p)^2} \tag{2}$$

$$D(x,p) = (x-p)^2 \tag{3}$$

$$D(x,p) = Abs(x-p) \tag{4}$$

$$D(x,p) = Max(|x-p|) \tag{5}$$

where $D(x, p)$ is the distance metric, x is the new point and p is the learning example. The regression or classification precision depends mainly on the metric used to calculate distances [18].

The calculations were performed in the programme Statistica 12.0. Operating data from the time span 1999–2014 (received from Water Utility) in one Polish water-pipe network were used for prediction purposes. The whole data set for years 1999-2013 was divided randomly into two groups (learning – 75% and testing – 25%). Besides, data from year 2014 were used for verifying the model. The verification data were not shown to the model previously. The failure rates of: water mains-λ_m, distribution pipes-λ_r and house connections-λ_p were forecasted value (dependent variable) on the basis of independent variables: number of installed house connections-LP, total length and the number of damages of water mains, distribution pipes and house connections-L_m, L_r, L_p and N_m, N_r, N_p, respectively. The short description of the water-pipe network and the city is as follows: The selected city is one of the oldest in Poland. It is located on the left bank of the Oder River in the western part of Poland. The town was founded in the 10th century. The first mentions of a water supply system being built in this city date from the middle of the 15th century [19]. At that time wells were dug from which water was supplied via conduits made of hollow pine wood to reservoirs. In the 16th century also pipes made of fired clay were used. In the next centuries the water supply system would be extended and rebuilt after numerous wars and fires. At the beginning of the 19th century some of the main conduits were replaced with cast-iron pipes. Towards the end of the 19th century water, in the amount of about 1050 m^3 per 24 h, would be supplied to consumers exclusively via cast-iron pipelines about 12 km long. Nowadays water supply system is fed with water from wells drilled from the Quaternary water-bearing horizon. After aeration, filtration and disinfection the water is pumped to the municipal system. Considered city has about 74000 inhabitants almost 100% of whom are connected to the system. The daily average supply is about 8300 m^3/d. Because of the topography and the height of the buildings, there are two water supply zones with pressures of 0.36 MPa and 0.56 MPa, respectively. The network includes: main conduits of both grey cast iron and steel transit, with diameters of DN 350 to 700 mm; distribution pipes with diameters of DN 80 to 250 mm, made variously of grey cast iron, steel, PE, PVC or AC; house connections with diameters of DN 25 to 200 mm, made of steel or PE [19]. The length of each kind of conduit is displayed in the Table 1.

3 Results and discussion

The ranges of experimental independent variables for years 1999–2014 are listed in the Table 1. The dependent variables varied from 0.07 to 0.57 fail./(km·a), 0.23-0.67 fail./(km·a) and 0.34-1.01 fail./(km·a) for water mains, distribution pipes and house connections, respectively. The considered failures occurred only on the pipes. The damages of fittings (e.g. on hydrants or valves) were not taken into consideration. We should distinguish the failure rate of linear objects (as pipes) and nonlinear objects (as fittings). The failure indicator should be calculated separately for these two kinds of water network elements. In this work the author would like to put emphasis only on the water conduits. All types of distance metrics were checked (E, E2, M and C). The results of prediction using E and E2 distance metrics were the same. As it was mentioned above the whole data set (1999–2013) was divided into learning (11 years) and testing (4 years) sample. The prediction results displayed in the tables 2–4 are related to testing sample. The learning sample was treated only as an example and the prediction was not carried out. The prediction results using verification sample (one year) are shown in the Figures 1–3.

Table 1 Independent variables

	LP	L_m, km	L_r, km	L_p, km	N_m	N_r	N_p
	1999–2013						
Min.	1954	28.2	99.2	35.4	2	27	15
Max.	2693	31.0	116.7	45.6	16	66	46
	2014						
	2721	31.0	118.4	46.3	7	33	42

The analysis of failure rate prediction of water mains (Table 2) indicates that models using Euclidean and quadratic Euclidean distance metrics are the most optimal. The convergence between experimental and predicted values of indicator λ is relatively good.

Table 2 Experimental and predicted failure rate of water mains (testing)

λ_m, fail./(km·a)			
Experimental	KNN-E and E2	KNN-M	KNN-C
0.32	0.32	0.27	0.22
0.14	0.18	0.14	0.13
0.14	0.13	0.14	0.13
0.17	0.13	0.13	0.13

On the other hand the best prediction of λ_m in verification step (Fig. 1) was obtained using Czebyszew distance metric. KNN models (relating to different distance metrics) were created all together for water mains, distribution pipes and house connections. In other words one model consisting of all independent variables LP, L_m, L_r, L_p and N_m, N_r, N_p was responsible for forecasting three dependent variables λ_m, λ_r and λ_p. Taking into consideration this approach, it seems to be reasonable to choose such model which would generate the smallest discrepancies between experimental and predicted values of failure rate relating to three types of conduits.

The quality of the prediction is also measured by the error of the validation. This error was equal to 0.0130, 0.0152 and 0.0150 for models KNN-E and KNN-E2, KNN-M and KNN-C, respectively. It means that models using Euclidean and quadratic Euclidean distance metrics seem to be the most optimal.

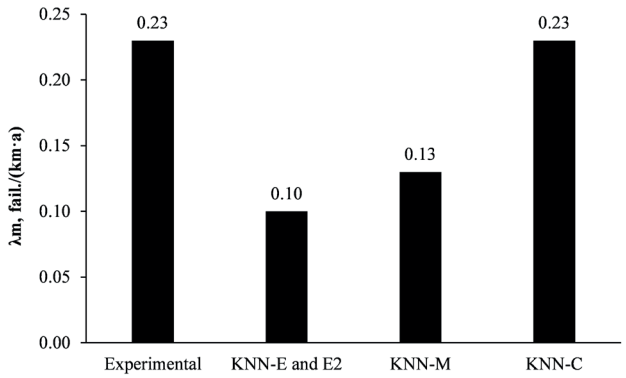

Fig. 1 Experimental and predicted failure rate of water mains (verification)

Table 3 Experimental and predicted failure rate of distribution pipes (testing)

λ_r, fail./(km·a)			
Experimental	KNN-E and E2	KNN-M	KNN-C
0.26	0.38	0.38	0.36
0.35	0.35	0.34	0.33
0.24	0.33	0.34	0.33
0.42	0.33	0.29	0.33

Concerning the failure rate prediction of distribution pipes it is obvious that not only in testing (Table 3), but also in verification (Fig. 2) KNN-E and KNN-E2 models are characterized by relatively good agreement between experimental and predicted values. For all types of conduits model KNN-C generated the constant value of indicator λ for three, from among four, testing years. It means that the model using Czebyszew distance metric is rather not recommended for forecasting purposes. Moreover, the distance is measured as a maximum of absolute value of differences between new and example points (equation (4)). Probably such kind of measurement of distance between points is not suitable for prediction of failure rate of water pipes.

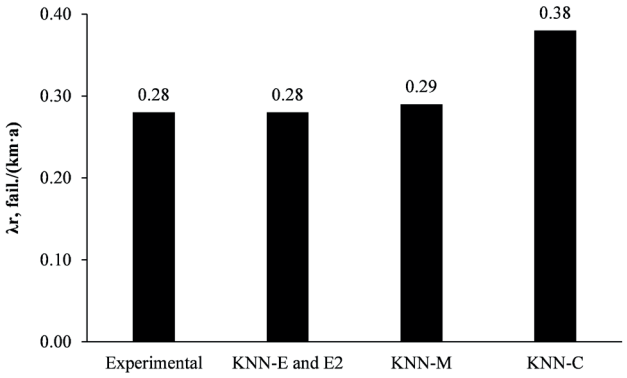

Fig. 2 Experimental and predicted failure rate of distribution pipes (verification)

From engineering point of view the prediction results (Table 4), using models KNN-E and KNN-E2, relating to house connections are satisfactory and could be accepted.

Table 4 Experimental and predicted failure rate of house connections (testing)

λ_p, fail./(km·a)			
Experimental	KNN-E and E2	KNN-M	KNN-C
0.76	0.75	0.70	0.65
0.53	0.47	0.43	0.35
0.36	0.35	0.43	0.35
0.64	0.35	0.57	0.35

The analysis of verification results (Fig. 3) indicates that model KNN-C forecasted indicator λ_p the most properly in comparison to other models. But taking into consideration the constraints and assumption that the model should be suitable for all types of conduits, also in testing step, model KNN-C does not seem to be the optimal one.

The graph of the changes of cross-validation error depending on the number of nearest neighbours is displayed in the Figure 4. The maximum number of nearest neighbours was estimated at the level of 11. This number was established by the algorithm in the programme Statistica 12.0.

Maximal number of nearest neighbours depends on the number of independent variables and number of cases. The analysis of the Figure 4 shows that the lowest validation errors were obtained when the optimal (minimum) number of K-nearest neighbours was equal to 2 and 3 for models KNN-E, KNN-E2, KNN-C and KNN-M, respectively. After the errors reached the lowest value, then the errors were increasing. We can assume that even if the number of nearest neighbours is higher, the errors will be still increasing.

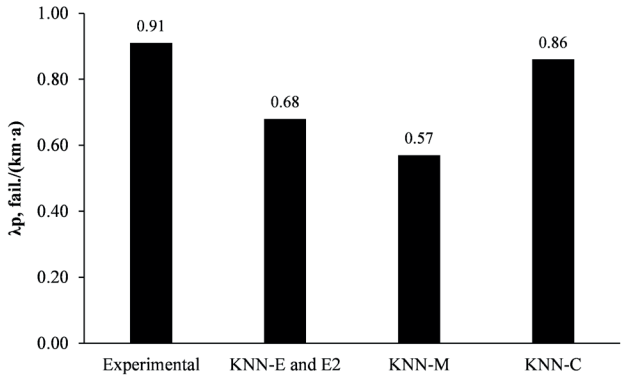

Fig. 3 Experimental and predicted failure rate of house connections (verification)

4 Conclusions

The K-nearest neighbours algorithm could be used for prediction of failure rate of water pipes. Analysis of the results obtained in the testing step shows that Euclidean and quadratic Euclidean distance metric seem to be the most appropriate for failure rate prediction for all types of pipelines. On the other hand, in the verification step Czebyszew and Euclidean distance metrics were the most suitable for failure rate forecasting of water mains, house connections and distribution pipes, respectively. Taking into account all assumptions and constraints, models KNN-E and KNN-E2 were supposed to be chosen as optimal. In this paper failure rate of water mains, distribution pipes and house connections were forecasted using independent variables related to three types of conduits. It means that each model consisted of three dependent variables (λ_m, λ_r, λ_p). Such assumption made at the very beginning forced to choose one model which was responsible for proper prediction of indicator λ of each type of water pipe. The results of forecasting and also the analysis of the lowest validation error pointed that the models characterized by Euclidean and quadratic Euclidean distance metrics were optimal. They also had the lowest number of K nearest neighbours which was equal to 2. It means that the models KNN-E and KNN-E2 were relatively simple that is quite important because among other aims the simplicity of the model is crucial issue in modelling.

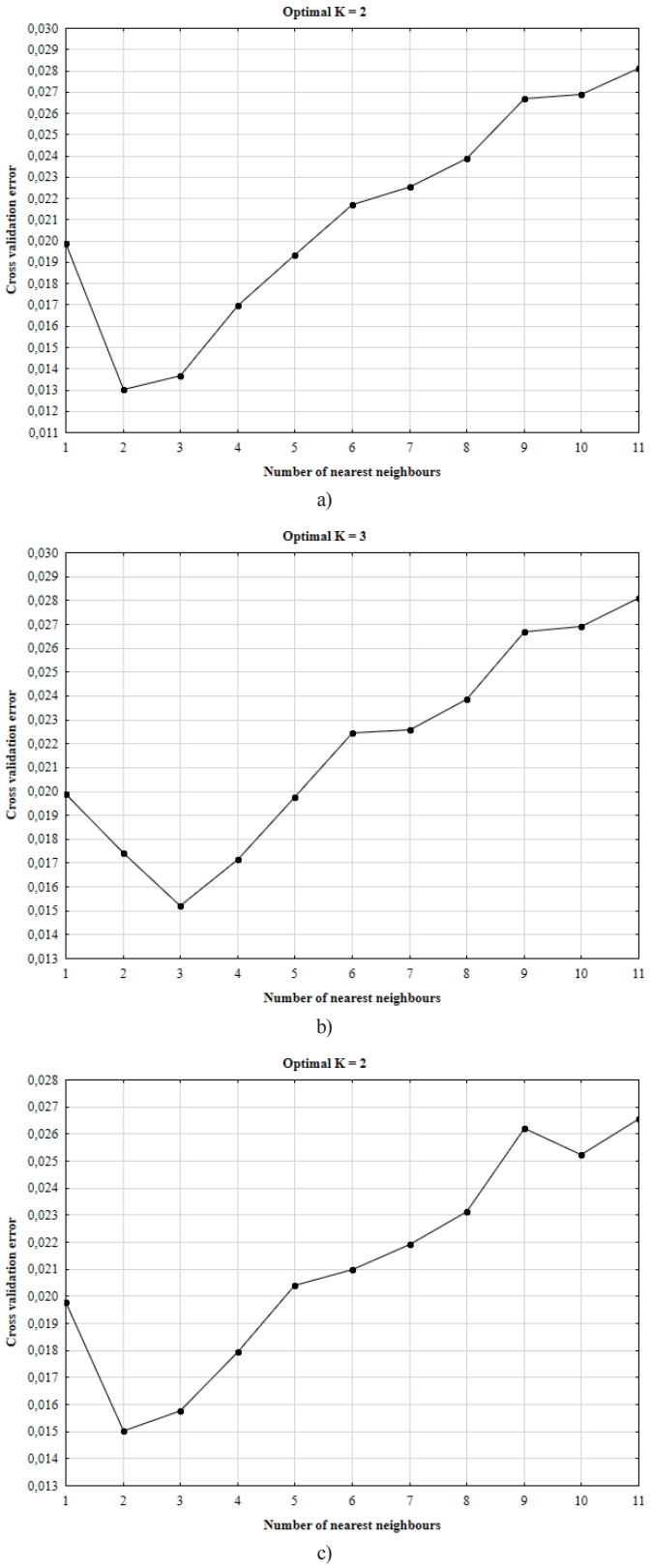

Fig. 4 Optimal number of nearest neighbours a) KNN-E and E2, b) KNN-M, c) KNN-C

The paper shows the very beginning step of failure rate modelling using K-nearest neighbours algorithm. The investigations presented in this paper will be expanded. The next step of researches would be checking whether other independent variables like e.g. material, diameter and year of installation of the water pipes are significant one for prediction purposes. The

models, for indicator λ separately for each type of conduit, will be also created and checked whether the convergence between experimental and forecasted values is better than concerning one model as it was presented in this paper.

References

[1] Cieżak, W., Cieżak, J. "Routine forecasting of the daily profiles of hourly water distribution in cities. An effectiveness analysis". *Environment Protection Engineering*, 41(2), pp. 179–186. 2015. https://doi.org/10.5277/epe150215

[2] Iwanek, M., Suchorab, P., Karpińska-Kiełbasa, M. "Suffosion holes as the result of a breakage of a buried water pipe". *Periodica Polytechnica Civil Engineering*, 61(4), pp. 700–705. 2017. https://doi.org/10.3311/PPci.9728

[3] Martinez-Codina, A., Castillo, M., Gonzales-Zeas, D., Garrote, L. "Pressure as a predictor of occurrence of pipe breaks in water distribution networks". *Urban Water Journal*, 13(7), pp. 676–686. 2016. https://doi.org/10.1080/1573062X.2015.1024687

[4] Hotloś, H. "Ilościowa ocean wpływu wybranych czynników na parametry i koszty eksploatacyjne sieci wodociągowych". (Quantitative assessment of the effect of some factors on the parameters and operating costs of water-pipe networks.) Wrocław University of Technology PublishingHouse, Wrocław, Poland, 2007. (in Polish) http://www.dbc.wroc.pl/Content/4273/Hotlos.pdf

[5] Bogardi, I., Fülöp, R. "A spatial probabilistic model of pipeline failures". *Periodica Polytechnica Civil Engineering*, 55(2), pp. 161–168. 2011. https://doi.org/10.3311/pp.ci.2011-2.08

[6] Pietrucha-Urbanik, K., Żelazko, A. "Approaches to assess water distribution failure". *Periodica Polytechnica Civil Engineering*, 61(3), pp. 632–639. 2017. https://doi.org/10.3311/PPci.10012

[7] Kleiner, Y., Rajani, B. "Comprehensive review of structural deterioration of water mains: statistical models". *Urban Water*, 3(3), pp. 131–150. 2001. https://doi.org/10.1016/S1462-0758(01)00033-4

[8] Renaud, E., Le Gat, Y., Poulton, M. "Using a break prediction model for drinking water networks asset management: From research to practice". *Water Science and Technology: Water Supply*, 12(5), pp. 674–682. 2012. https://doi.org/10.2166/ws.2012.040

[9] Kutyłowska, M. "Neural network approach for failure rate prediction". *Engineering Failure Analysis*, 47, Part A, pp.41–48. 2015. https://doi.org/10.1016/j.engfailanal.2014.10.007

[10] Nishiyama, M., Filion, Y. "Forecasting breaks in cast iron water mains in the city of Kingston with an artificial neural network model". *Canadian Journal of Civil Engineering*, 41(10), pp. 918–923. 2014. https://doi.org/10.1139/cjce-2014-0114

[11] Scheidegger, A., Leitao, J. P., Scholten, L. "Statistical failure models for water distribution pipes – A review from a unified perspective". *Water Research*, 83. pp. 237–247. 2015. https://doi.org/10.1016/j.watres.2015.06.027

[12] Wang, Y., Zayed, T., Moselhi, O. "Prediction models for annual break rates of water mains". *Journal of Performance of Constructed facilities*, 23(1), pp. 47–54. 2009. https://doi.org/10.1061/(ASCE)0887-3828(2009)23:1(47)

[13] Kaźmierczak, B., Wdowikowski, M. "Maximum rainfall model based on archival pluviographic records – case study for Legnica (Poland)". *Periodica Polytechnica Civil Engineering*, 60(2), pp. 305–312. 2016. https://doi.org/10.3311/PPci.8341

[14] Candelieri, A., Soldi, D., Conti, D., Archetti, F. "Analytical leakages localization in water distribution networks through spectral clustering and support vector machines. The Icewater approach". *Procedia Engineering*, 89, pp. 1080–1088. 2014. https://doi.org/10.1016/j.proeng.2014.11.228

[15] Malinowska, A. "Classification and regression tree theory application for assessment of building damage caused by surface deformation". *Natural Hazards*, 73(2), pp. 317–334. 2014. https://doi.org/10.1007/s11069-014-1070-2

[16] Illa, J. M. G., Alonso, J. B., Marre M. S. "Nearest-Neighbours for time series". *Applied Intelligence*, 20(1), pp. 21–35. 2004. https://doi.org/10.1023/B:APIN.0000011139.94055.7a

[17] Statistica 12.0, Electronic Manual.

[18] Weinberger, K. Q., Saul, L. K. "Distance metric learning for large margin nearest neighbour classification". *Journal of Machine Learning Research*, 10, pp. 207–244. 2009. http://jmlr.csail.mit.edu/papers/volume10/weinberger09a/weinberger09a.pdf

[19] Kutyłowska, M., Hotloś, H. "Failure analysis of water supply system in the Polish cityof Głogow". *Engineering Failure Analysis*, 41, pp. 23–29. 2014. https://doi.org/10.1016/j.engfailanal.2013.07.019